A PRESCIENT ABSTRACT:

In the past 30 years, various ideas have been presented for machines which could recognize spatial patterns. With the advent of the digital computer and its use in data processing, there has been a great increase of interest in the automatic conversion of human language to language understandable by a machine. The translation to machine language of spatial symbols---pattern recognition---is important to this conversion.

Papers presented at the the March 3-5, 1959, Western Joint Computer Conference San Francisco, California  Pages: 291-294
W. H. Highleyman (RPI 1955),  Bell Telephone Labs., Inc., Murray Hill, N.J.
L. A. Kamentsky  Bell Telephone Labs., Inc., Murray Hill, N.J.
Document Systems Analysis: Testing, Testing, Testing...

A Short History of Document Test Data

by

George Nagy

DocLab, RPI

PG-50
Organizations & Committees ~ 1967

IEEE-CG TC on Pattern Recognition

Subcommittee on Reference Data Sets
First Pattern Recognition Meeting
Oct. 1966 - chaired by Al Hoagland

Abend, Ball, Chow, Cover, Duda, Freeman, Groner, Hall, Julesz, Kanal, Kirch, Minsky, McCormick, Munson, Prewitt, Roberts, Papert, Rabinow, Roberts, Rosenfeld, Sammon, Specht, Stanat, Sheinberg, Sutherland, Specht, Uhr, Widrow, Zadeh...
A HAPPENING IN PUERTO RICO

Trends in Pattern Recognition
A Report on the 1966 IEEE Pattern Recognition Workshop

by
George Nagy
IBM Watson Research Center
Yorktown Heights, New York

Who, What, When, Where

Self-Organizing, Bionic, Heuristically Programmed, Pattern Recognizing, Learning, Neuronal, Cybernetic, Goal-Seeking, Problem-Solving, Microprogrammed, Multiprogrammed, Multi-Input, Redundant, Adaptive, Self-Repairing, Self-Teaching, Time-Sharing, Self-Reproducing, Cluster-Seeking, On-Line, Trainable, Stochastic, Kilometreshdy, Optimal, Artificially Intelligent, Symiotic Computing Machines — was one speaker's list of the key words necessary to describe the range of topics discussed at the recent "happening" (the chairman's characterization) instigated in Puerto Rico by the Pattern Recognition Subcommittee of the IEEE Computer Group.

the user to make on-the-spot corrections and to adjust style to suit the recognition logic. The one-off-line procedure from relies on the context inherent in a programming language such as Fortran to keep the error rate within acceptable limits.

Further contributions in character recognition consisted new algorithms designed to improve maximum likelihood decisions based on features by taking into account the int feature statistical dependencies and the Markovian property of natural language.

Other applications-oriented presentations covered ho graphic techniques for fingerprint recognition, polymorphic decision boundaries for electrocardiograms, automated photo metric chromosome analysis, adaptive networks for phased antenna arrays and for aerial photoreconnaissance, a sequential decision model for blackjack, graphic input computers, the superposition of flight paths on contour maps, and the analysis of three-dimensional projections. And these, the ECG analysis seems closest to practical applicability. Several of the other projects, notably the work on fingerprints, chromosomes, sonar, and graphic input, also make use of realistic data sets.

The outline of a general purpose pattern recognition a
Minutes of the
IEEE Computer Group
Technical Committee
on Pattern Recognition (PRC) and its
Subcommittee on Reference Data Sets and
Performance Evaluation Standards (SCDS)

- PR Subcommittee upgraded to full TC
- Puerto Rico Workshop proceedings:
  - Wiley would typeset and publish for $10 per copy
  - Spartan would print “as is” for $5-6 per copy
  - McGraw Hill not interested, Addison Wesley maybe
  - 8000 copies of OPTICAL CHARACTER RECOGNITION (Spartan 1960) sold
  - IEEE SMC PG-35 interested (also in taking over PG -16)
- Pattern Recognition Society established by R. Ledley
April 10, 1967, continued

• Participation at forthcoming meetings:
  – Washington, DC (UMD, ONR, USPS) June 1967 - Rosenfeld
  – Systems Science Hawaii January 1968 – Watanabe
    Specifically, Nagy’s controversial short paper on Unsupervised Learning,
    presented at the Puerto Rico worship, should be expanded to a full paper (???)
  – IFIPS Edinburgh August 1968 – Herb Freeman
    – IEEE G-SSC TC on Pattern Recognition and Learning Systems
    – IEEE G-AC TC on Adaptive and Learning Systems

(MMS+SSC → SMC)

After the meeting, Al Hoagland asked me to look into reference data.
MEMORANDUM

To: A. S. Hoagland, Chairman, IEEE Pat Rec Committee
From: G. Nagy
Date: August 28, 1967
Subject: **Reference Data Sets**

Contacted universities, OCR manufacturers, industrial and non-profit research labs, and the United States Government.

“The consensus of opinion appears to be that the establishment of reference material of this sort is long overdue.”

Alternatives:

1. Establish a small subcommittee to publish a **catalog** of data sets which may be obtained from their respective owners;

2. Co-sponsor with a **government agency** (NIST, then NBS);

3. Endorse a **non-profit research org** or a **consultant** to make reference data available to the public for a fee;

4. A means within IEEE to acquire, maintain and distribute the data, in a manner similar to the Computer Group Repository, which in fact already encourages the submission of original data.
• Procs. of Puerto Workshop mailed to Thompson (Kanal).
• Nagy’s suggestion for a subcommittee on reference data sets adopted. Requests to add pictorial and time-varying signals, as well as multi-dimensional feature data.

[CK Chow, Munson (SRI), Spooner (CAL), Samson (RDC), Gibbons (PO)].

• Duda designated Committee’s Contributing Editor
  – (to Computer Group News).
• PRS is establishing a new journal: Pattern Recognition
  (editors from PRC invited).
• Explore cooperation in IP with Optical Society of America.
• Delft PR workshop in August 1968 (Verhagen).
• Pisa two-week summer school in September 1968 (Grasselli).
• San Francisco SSC Conf in October 1968.
SCDS March 21, 1968 NYC

- C.K. Chow received enthusiastic IEEE endorsement and offer to cooperate.
- Harry Huskey will accept data tapes as items of his (PGEC) repository: *IEEE Test Data Sets in Computer Group News, March 1968, p. 28.*
- Harburgen obtained legal advice: IEEE should copyright data, identifying and giving full credit to source. IEEE cannot restrict use of data to researchers.
- MP: Spooner has found several MP sources. REI has large data base, but reluctant to release it except to researchers. C.K. Chow said IBM may release 10,000 chars of MP. CAL has 40,000 alphanumeric chars in 24x24 binary form with 64 levels. Expect soon to have 150,000 chars including IBM Executive and Remington Rand Elite.
- HP: Highleyman; Knoll; Munson
- Cursive: Harmon, Murray Eden, Gibbons
- Speech; EKG, EEG, EMG; Seismic; Radar/Sonar; Fingerprints; X-Rays; Bubble Chamber, Microscope, Aerial, Celestial Photos; Maps; Line Dwgs (none!); Property lists (Medical diagnostics, Taxonomy);
  W.H. Highleyman
• 203 copies of Pattern Recognition sold!
• Should IEEE repository handle data set distribution (efficiency and cost)?
• No formal efforts will be made to create an umbrella organization to coordinate pattern recognition activities among IEEE committees and other groups (the autonomy of the Computer Group with the IEEE is now under discussion.)
• Delft Workshop report by Chandrasekaran, Kanal and Nagy published by the IEEE.
• Hoagland thanked and sent up, Chow welcomed as TC chair.
• Workshop in Honolulu proposed.

The problem of verifying the description of data on magnetic tape was seen to be a formidable one because of incompatibilities among different computer systems.

Hamburgen will check IEEE Computer Center facilities.
• Four documented data sets submitted to IEEE HQ. Requires IEEE ADCOM and Tech Ad Board approval: Chow wrote to McCluskey

• Twelve facilitators plus two international coordinators appointed for data bases

• Subcommittee renamed: Reference Data Bases and Evaluation Procedures

• Hawaii workshop with SMC should have no session on OCR or statistical techniques, sessions should not be organized by subject, and most sessions should be devoted to future methods of pattern recognition. (Objection: this leaves out most engineers who work in pattern recognition)
Spooner to chair SCRD
Hamburgen will organize a subcommittee on character recognition
Still 672 copies left of Pattern Recognition!
Still looking for a Service Bureau to copy tapes
Preparations for 1972 Workshop on PR in Hot Springs (Ed Parrish)
1. Data Sets sent from IEEE HQ to Computer Society two months ago, but have not yet arrived

5. Should advertise in *IEEE Spectrum*

6. Publish names of data set users

7. Rabinow to be asked to obtain a set of alphanumeric characters that have been systematically degraded

9. “It was first agreed that the first order of business was to get the data sets in hand and advertise in Computer so that we can point to a significant milestone in the Committee’s accomplishments.”
Five data sets are ready and will be advertised in the January-February 1972 Issue of *Computer* magazine.

Production cost estimated at $100.

PRC should consider expanding its scope to *image processing* and *artificial intelligence*.

Al Klinger requested to prepare a Glossary of Pattern Recognition.
Data Sets advertised in IEEE Computer

January 1972 (6 data sets)

In order to encourage research in the field of pattern recognition, the IEEE Computer Society’s Technical Committee on Pattern Recognition has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the bottom of the page.

1.1.1 Machine Imprinted Alphanumeric Characters – Dr. H. F. Ryan, Calcomp Corp. U.S. Postal Service
An alphanumeric character data base of 100,000 samples of 66 character classes. (Also known as the CAL-U.S. Postal Service Alphanumeric Character Data Base.) Thresholded binary images of segmented, centered, midpoint, machine-imprinted characters. Resolution is 24 x 24. Magnetic tape, 9 track, 2 reels, 1600 BPI.
Price: $112.50 ($98.75 with furnished tape)
Member’s discount price: $90, ($66.65 with furnished tape)

1.2.1 Handprinted Numeral Characters – Dr. A. L. Knoll, Honeywell Information Systems, Data Systems Division
The data base consists of 50 samples of each numeral character. Simple printing rules were specified but not always followed. The samples were selected from those contributed. The images are binary with a resolution of 28 x 21. Punch cards.
Price: $41.25
Member’s discount price: $33.

1.2.2 Handprinted FORTRAN Alphanumeric Characters – Dr. John H. Muson, Stanford Research Institute
The data base consists of two parts, each part on a reel. The first part contains 3 alphabets of 46 characters, corresponding to the nine-black character set of the basic FORTRAN language. Hand-printed by each of 49 writers, making a total of 3 x 46 x 49 = 6,762 patterns.
The second part has 2,999 characters printed by a single writer. There are 920 characters made up of 20 alphabets of 46 characters each; the remaining 2,079 characters are taken from fragments of actual coding sheets. The images are binary with a resolution of 24 x 24 resolution. Magnetic tape, 7 track, 2 reels, 556 BPI.
Price: $196.75 ($173.50 with furnished tape)

Pattern Recognition

DATA BASES AVAILABLE

In order to encourage research in the field of pattern recognition, the IEEE Computer Society’s Technical Committee on Pattern Recognition has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the bottom of the page.

1.2.4 Handprinted Numeric Characters – Hiroshi Genclo, Asia City Shibaura Electric Co., Ltd/Nishindai Research and Development Center
The data base consists of 10,000 hand-written numeric characters collected from five as well as experimental mail throughout Japan. The data base is contained on two magnetic tapes with each tape having 2,000 characters. Images are binary with a resolution of 36 x 36. A single pattern consists of 36 x 36 individual characters. Magnetic tape, 7 track, 2 reels, 556 BPI, 6 bits per character.
Price: $115.75 ($106.25 with furnished tape)
Member’s discount price: $92, ($86, with furnished tape)

1.3.1 Cursive Script – Dr. L. D. Harmon, Bell Telephone Laboratories

In order to encourage research in the field of pattern recognition, the IEEE Computer Society’s Technical Committee on Pattern Recognition has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the bottom of the page.

1.1.1 Machine Imprinted Alphanumeric Characters – Dr. H. F. Ryan, Calcomp Corp.
An alphanumeric character data base (normalized version) of 100,000 samples of 66 character classes. Thresholded binary images of segmented, centered, midpoint, machine-imprinted characters. Resolution is 24 x 24. Magnetic tape, 9 track, 2 reels, 1600 BPI.
Price: $123.75 ($107.50 with furnished tape)
Member’s discount price: $99 ($86 with furnished tape)

Pattern Recognition

DATA BASES

1.1.1A Machine Imprinted Alphanumeric Characters – Dr. H. F. Ryan, Calcomp Corp.

To order: Use the multipurpose form at the back of the issue.

April 1973

January 1973 (6 data sets)
Announcement

In order to encourage the field of pattern recognition, the IEEE Computer Society's Technical Committee on Machine Pattern Analysis has begun collecting data bases from a variety of sources. These data bases, including substantial back-up documentation, may be ordered by using the form at the back of the issue.

Discounts off the data base list prices are available to IEEE members and members of the American Federation of Information Processing Societies' constituent societies.

When ordering, you may elect to send us your own blank tapes; if you do, be sure they are in good condition and have no other data recorded on them.

If you have a data base that you wish to contribute to the Technical Committee on Machine Pattern Analysis, please contact Dr. J. B. McFerran, Sperry-Univac, P. O. Box 3525, St. Paul, Minnesota 55165.
PRC  June 17, 1972  Atlantic City

• 20 Data Sets purchased!
• Ed Parrish to chair SCDS
• 41 @ Hot Spring PR Workshop. Income – Expenses = $705.92.
• S. Yau: Expand scope of PRC and change name, or form another committee:

Pattern Recognition                           Complex Information Processing
Pattern Recognition and Machine Intelligence  Heuristic Problems
Pattern Recognition and Artificial Intelligence Cognitive Technology
Models of Cognition (or Intelligence)          Artificial Intelligence
Machine Pattern Analysis

SCDS

Suggestion: find a “willing expert “ to verify data sets.
McFerran explained Univac’s computer-generated character data bases.
> 100 attendees at 2-DSP Conf at U. Missouri in August 1972

37 Data Sets purchased!

- Machine imprinted alphanumeric characters 1.1.1 7
- Handprinted numeric characters 1.2.1 7
- Handprinted FORTRAN alphanumeric characters 1.2.2 10
- Handprinted alphanumeric characters 1.2.3 7
- Handprinted numeric characters 1.2.4 3
- Cursive script 1.3.1 3

37

PRC June 7, 1973 NYC

60 Data Sets purchased!

Will be advertised regularly in *Computer*

Should software (programs) be added?
Lengthy discussion of acceptability of artificially generated data. Collaborate with ACM SIGART?

PRC May 7, 1974 Palmer House
Task force on publishing mechanisms in PR-oriented journals

PRC Sept 11, 1974 Mayflower
Copenhagen
Silver Spring
Asilomar
El Coronado San Diego

IJCSR June 1974: 400 attendees
IPR Workshop November 1974
PR Workshop March 1975
ICPR Fall 1976

First Machine Pattern Analysis TC Newsletter October 1974
1974 PRC

Agrawala, Brick, Butler, Butterfield, Chien, Chow, Deutsch, Fischler, Frank, Freeman, Fukunaga, Gibbons, Hamburgen, Harlow, Hoagland, Kanal, Klinger, Lainiotis, Lambert, Lederer, McFerran, Mathur, Meisel, Nadler, Nagy, Parrish, Patrick, Robinson, Rosenfeld, Samit, Sammon, Shapiro, Sklansky, Spooner, Stoffel, Swonger, Watanabe, Weinstein, Wilson, Widrow, Yau
Most often used dataset ever (50 alphabets)

1.2.3 Handprinted Alphanumeric Characters – Dr. W. H. Highleyman, Sombers Associates.

There are approximately 2300 samples of alphanumeric characters. The images are binary with a resolution of 12 x 12. Punched cards.

Price: $55
Member’s discount price: $44

over 50 requests by August 1967

IEEE Computer, April 1976, p. 83
1.2.2 Handprinted FORTRAN Alphanumeric Characters
Dr. John H. Munson, Stanford Research Institute

The data base consists of two parts, with each part on a reel. The first part contains 3 alphabets of 46 characters, corresponding to the non-blank character set of the basic FORTRAN language, hand-printed by each of 49 authors making a total of $3 \times 46 \times 49 = 6,762$ patterns.

The second part has 2,999 characters printed by a single author. There are 920 characters made up of 20 alphabets of 46 characters each; the remaining 2,079 characters are taken from fragments of actual coding sheets. The images are binary with a 24 x 24 resolution. Magnetic tape, 7 track, 2 reels, 556 BPI.

Price: $116.75 ($75.75 with furnished tapes)
Member's discount price: $93 ($60 with furnished tapes)
Fig. 14. Handprinted characters on FORTRAN coding sheet. An augmented FORTRAN alphabet is shown by each of twelve different writers. The range of variation is considerable even though the writers were in no particular hurry. These data were collected at the Stanford Research Institute under sponsorship of the U. S. Army Electronics Command.

<table>
<thead>
<tr>
<th>COLS. I-10</th>
<th>COLS. II-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9</td>
<td>A, B, C, D, E</td>
</tr>
<tr>
<td>U,V,W,X,Y,Z</td>
<td>=, +, ..., $</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COLS. I-10</th>
<th>COLS. II-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9</td>
<td>A, B, C, D, E</td>
</tr>
<tr>
<td>U,V,W,X,Y,Z</td>
<td>=, +, ..., $</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COLS. I-10</th>
<th>COLS. II-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9</td>
<td>A, B, C, D, E</td>
</tr>
<tr>
<td>U,V,W,X,Y,Z</td>
<td>=, +, ..., $</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COLS. I-10</th>
<th>COLS. II-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9</td>
<td>A, B, C, D, E</td>
</tr>
<tr>
<td>U,V,W,X,Y,Z</td>
<td>=, +, ..., $</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COLS. I-10</th>
<th>COLS. II-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9</td>
<td>A, B, C, D, E</td>
</tr>
<tr>
<td>U,V,W,X,Y,Z</td>
<td>=, +, ..., $</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COLS. I-10</th>
<th>COLS. II-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9</td>
<td>A, B, C, D, E</td>
</tr>
<tr>
<td>U,V,W,X,Y,Z</td>
<td>=, +, ..., $</td>
</tr>
</tbody>
</table>
1.1.2 Machine Imprinted Multilevel Characters
- James S. Gibbons, Electronic Sciences Division, U.S. Postal Service.

An alphanumeric character data base of 32,000 (16 level) multi-level characters, extracted from 2100 test mail pieces each in a 24 x 24 array covering an area of 0.144 x 0.144 square inches. Complete information as to upper/lower case, font style, percent reflectance, color print and background accompanies each character on magnetic tape. Suitable for machine independent research as well as OCR testing. Magnetic tape, 4 reels test characters, 1 reel set up characters, 9 track, 800 BPI.

Price: $262.50 ($150.00 with furnished tapes)
Member's discount price: $210.00 ($120.00 with furnished tapes)

11.2.1 Handprinted Numeric Characters -
Dr. A. L. Knoll, Honeywell Information Systems, Data Systems Division

The data base consists of 50 samples of each numeric character generated by 9 different authors. Simple printing rules were specified but not always followed. The samples were selected from those contributed. The images are binary with a resolution of 25 x 21. Punched cards.

Price: $41.25
Member's discount price: $33
1.2.4 Handprinted Numeric Characters - Hiroshi Genchi, Tokyo Shibaura Electric Co., Ltd./Toshiba Research and Development Center
The data base consists of 10,000 hand written numeric characters collected from live as well as experimental mail throughout Japan. The data base is contained on two magnetic tapes with each tape having 5,000 characters. Images are binary with a resolution of 36 x 50. A single pattern consists of 56 words. Magnetic tape, 7 track, 2 reels, 800 BPI, 6 bits per character.

Price: $115.25 ($68.75 with furnished tapes)
Member's discount price: $93 ($55 with furnished tapes)

1.3.1 Cursive Script - Dr. L. D. Harmon, Bell Telephone Laboratories
The data consists of 52 cursive script sentences. The resolution for each sentence is 256 (vertically) x 2048 (horizontally). The images are binary. Magnetic tape, 7 track, 1 reel, 200 BPI.

Price: $60.50 ($42.25 with furnished tapes)
Member's discount price: $50 ($33 with furnished tape)

To order: Use the multipurpose order form at the back of the issue.
OCR infrastructure: Scanners

Drum scanners for telegraphy

Bain 1841

Bakewell 1847

Caselli 1861

Nipkow disk

Rotating mirror

CRT

June 9, 2010

DAS 2010 (GN)
OCR GENEOLGY <1960

GUSTAV TAUSEK PATENT 1929
(MECH. TEMPLATES + PHOTODETECTOR)

DAVID SHEPARD 1951
INTELLIGENT MACHINES RESEARCH
READERS DIGEST, STANDARD OIL 1955
IMR → FARRINGTON → COGNITRONICS

JACOB RABINOW 1952
NATIONAL BUREAU OF STANDARDS 1952
US & CANADA POST 1953
US AIR FORCE (BORROUGHS) 1959
OCR Genealogy continued

1960  Farrington  Philco-Ford  NDP  Sperry  Univac

1962  Rabinow  Engineering  HP \(\rightarrow\) Control Data

> 1960  Fujitsu  Hitachi  NCR  IBM  RCA  GE  Solatron  Scan-Data
REI  ...  > 50 OCR companies in the United States

(40 docs & 10,000 chars per second, $, $$$, $$$$

> 1980  Kurzweil \(\rightarrow\) Xerox \(\rightarrow\) Scansoft \(\rightarrow\) Nuance

\[\uparrow\]

Palantir \(\rightarrow\) Calera \(\rightarrow\) Caere  \(\leftarrow\) Recognita

THOCR  Expervision  Fuji  Sanyo  RAF  IRIS  ABBYY...

> 2000  HP \(\rightarrow\) Tesseract \(\rightarrow\) OCRopus  GOOCR (GNU)
1962 Ray Bonner: A “Logical Pattern” Recognition Program

27,519 unique 7x10 bitmaps culled from 1,000,000 samples

2 substitution errors and 14 rejects
1966: IBM-1975 Social Security Page Reader

- 500 lines per minute; 200 fonts
- Last Quarter, 1965:
  - 1,300,000 pages = 33 million lines
- 16 million lines recognized
- 4 million lines with one or more rejects
- 12 million lines on rejected pages
- OCR confusion probabilities + SSA master file of 150 million names
  (one million distinct names with frequencies)
  (no cross-checking between names and SSN's)
<table>
<thead>
<tr>
<th>Employee ID</th>
<th>Name</th>
<th>Wages ($)</th>
<th>Total ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>156 89 3051</td>
<td>Mable Day</td>
<td>900.00</td>
<td></td>
</tr>
<tr>
<td>860 91 2191</td>
<td>Henry Long</td>
<td>45.50</td>
<td></td>
</tr>
<tr>
<td>985 10 3050</td>
<td>John Doe</td>
<td>4.95</td>
<td></td>
</tr>
<tr>
<td>110 06 4792</td>
<td>Pete Oak</td>
<td>372.90</td>
<td></td>
</tr>
<tr>
<td>589 36 2875</td>
<td>Jo Ellen Sand</td>
<td>650.00</td>
<td></td>
</tr>
<tr>
<td>799 29 6869</td>
<td>Chad Wood</td>
<td>90.35</td>
<td></td>
</tr>
<tr>
<td>100 45 5324</td>
<td>Aden Harmes</td>
<td>470.39</td>
<td></td>
</tr>
<tr>
<td>489 32 5031</td>
<td>Ethel Wilson</td>
<td>74.95</td>
<td></td>
</tr>
<tr>
<td>001 78 2543</td>
<td>Ardean Nelson</td>
<td>51.50</td>
<td></td>
</tr>
<tr>
<td>992 64 3065</td>
<td>Irving Jones</td>
<td>49.00</td>
<td></td>
</tr>
<tr>
<td>821 75 2193</td>
<td>Gertrude Swanson</td>
<td>725.00</td>
<td></td>
</tr>
<tr>
<td>207 91 5882</td>
<td>Barbara Smith</td>
<td>45.00</td>
<td></td>
</tr>
<tr>
<td>340 12 1478</td>
<td>Anna Evans</td>
<td>725.00</td>
<td></td>
</tr>
<tr>
<td>475 26 7282</td>
<td>Carl Hanson</td>
<td>529.65</td>
<td></td>
</tr>
<tr>
<td>342 12 3331</td>
<td>Herman Tree</td>
<td>370.73</td>
<td></td>
</tr>
<tr>
<td>635 56 1381</td>
<td>Mike Case</td>
<td>2.50</td>
<td></td>
</tr>
<tr>
<td>774 37 4593</td>
<td>John Parken</td>
<td>673.00</td>
<td></td>
</tr>
<tr>
<td>942 21 7833</td>
<td>Tom Hermanns</td>
<td>74.00</td>
<td></td>
</tr>
<tr>
<td>229 33 4381</td>
<td>Eugene Toogood</td>
<td>839.77</td>
<td></td>
</tr>
<tr>
<td>955 42 1872</td>
<td>Hilma Paul</td>
<td>59.63</td>
<td></td>
</tr>
<tr>
<td>680 11 3582</td>
<td>Ardella Hermann</td>
<td>730.00</td>
<td></td>
</tr>
<tr>
<td>524 73 7532</td>
<td>Greta Olson</td>
<td>89.65</td>
<td></td>
</tr>
<tr>
<td>774 26 4255</td>
<td>Betty Head</td>
<td>112.50</td>
<td></td>
</tr>
<tr>
<td>436 28 2743</td>
<td>Pat Mousebrown</td>
<td>632.75</td>
<td></td>
</tr>
<tr>
<td>357 89 0367</td>
<td>Sharon Ness</td>
<td>643.25</td>
<td></td>
</tr>
</tbody>
</table>

**NOTES**

- **TOTAL WAGES**: $10,232.46
- **Number of Employees**: 13
Reference data sets in the modern era
In order to develop and evaluate recognition methods, a common character database is necessary. The authors collected a database of handprinted Chinese characters on a magnetic tape. The database consists of 48,000 characters (4,000 categories $\times$ 12 sets). Each character is quantized to 1 bit and is sampled as 64 (horizontal) $\times$ 63 (vertical) frames. In this paper, the database is introduced and its fundamental topological features are evaluated. In particular, the number of connected components and that of holes are compared with features obtained from the handprinted Japanese Educational Kanji character database (ETL-8).

1. INTRODUCTION
1982 ETL-9(B)


200 samples each of first level JIS (71 Hiragana + 2965 Kanji)

32 x 32 bitmaps 607,200 characters
## Fin-de-siècle sources of western test data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBS-NIST</td>
<td>1975</td>
<td>300K HP alphanumeric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>91K phrases of real census data</td>
</tr>
<tr>
<td>CENPARMI</td>
<td>1981</td>
<td>100K HP 64x32 alphanumeric</td>
</tr>
<tr>
<td>ISRI</td>
<td>1993 - 96</td>
<td>Zoned page images</td>
</tr>
<tr>
<td></td>
<td>1992</td>
<td><em>On-line</em> cursive words</td>
</tr>
<tr>
<td>CEDAR</td>
<td>1994</td>
<td>15K HW “words” from mail</td>
</tr>
<tr>
<td>IAM</td>
<td>1999</td>
<td>HW sentences</td>
</tr>
</tbody>
</table>

*cf. Guyon, Haralick, Hull, Phillips, Data Sets for OCR and Document Image Understanding Research, Chapter 30, Handbook of Character Recognition and Image Analysis, Bunke & Wang, 1997*
## List of Available ISRI Test Datasets
(Nartker, Rice, and Lumos *SPIE/IS&T 2005*)

<table>
<thead>
<tr>
<th>Ground-truth Test Datasets</th>
<th>Number of</th>
<th>Image Resolution</th>
<th>Used in Annual Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Name</td>
<td>Description</td>
<td>Pages</td>
<td>Characters</td>
</tr>
<tr>
<td>Sample 2</td>
<td>DOE Sample 2</td>
<td>460</td>
<td>817,946</td>
</tr>
<tr>
<td>Sample M</td>
<td>Magazine Sample</td>
<td>200</td>
<td>666,134</td>
</tr>
<tr>
<td>Sample N</td>
<td>Newspaper Sample</td>
<td>200</td>
<td>492,080</td>
</tr>
<tr>
<td>Sample B</td>
<td>Business Letter Sample</td>
<td>200</td>
<td>319,756</td>
</tr>
<tr>
<td>Sample L</td>
<td>Legal Document Sample</td>
<td>300</td>
<td>372,098</td>
</tr>
<tr>
<td>Sample S</td>
<td>Spanish Newspaper Sample</td>
<td>144</td>
<td>348,091</td>
</tr>
<tr>
<td>Sample 3</td>
<td>DOE Sample 3</td>
<td>785</td>
<td>1,463,512</td>
</tr>
<tr>
<td>Sample R</td>
<td>Annual Report Sample</td>
<td>300</td>
<td>892,266</td>
</tr>
<tr>
<td>Sample Z</td>
<td>Magazine Sample 2</td>
<td>300</td>
<td>1,244,171</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td></td>
<td><strong>2889</strong></td>
<td><strong>6,616,054</strong></td>
</tr>
</tbody>
</table>
NIST OCR DATABASES

NIST Scoring Package
NIST Structured Forms Reference Set of Binary Images
NIST Structured Forms Reference Set of Binary Images II
NIST Machine-Print Database of Gray Scale and Binary Images
NIST Miniform Training Database
NIST Miniform Training Database II
NIST Miniform Test Database
NIST Handprinted Forms and Characters
NIST Scientific and Technical Document
NIST Federal Register Document Image Database
A new project at NIST (cosponsored by DOD) intended to bring developers of Optical Character Recognition (OCR) and Information Retrieval (IR) technologies together to study and evaluate the automated production and usability of large-scale, on-line collections of digital documents.

1. How should these collections be constructed in an automated way using OCR technology?
2. What impact do OCR errors have on the accuracy of IR?
3. What types of information should be provided from OCR processing (in addition to text) that will facilitate and enhance IR?
4. What types of data (in addition to text) should be indexed and retrieved to improve the usability of these collections?
5. What types of "real-world" OCR/IR integrated applications can be solved today, solved next year, solved in five years?

To be based on 68,000 scanned pages of a year’s worth of the Federal Register plus GPO typesetting files (SGML)
What is METADATA?

Foil 3 Mike Garris 1997

- Non-text elements (physical or logical) of a document
  - Fonts, page layout, equations, figures, tables, document type, language, title, author, dates, ...

- Metadata may be used to construct queries and/or it may be part of the information retrieved.

- Working definition of metadata will be developed by the planning committee.
PROPOSED SIZE OF DATA SETS

Training Set: 2,000 FR94 page images
2,000 FR94 ground truth files
100 of the 2,000 designated for evaluation
5 known-item queries

Testing Set: 10,000 FR94 page images
10,000 FR94 ground truth files (IR participants only)
100 known-item queries
Our primary interest in METTREC has been to pursue the use of automatically recognized metadata and measure its impact on information retrieval.

Based on our experience over the past year, we have observed:

a.) An organized "OCR research community" no longer exists. There only remains a small number of commercial vendors competing in a "shrinking" market.

b.) Very little research has been developed into technology tools for automatically detecting metadata in legacy paper documents that can be used for IR.
c.) There is little motivation for OCR participation.
d.) No one in the IR community is actively researching the use of metadata. It is acknowledged that metadata is interesting and might be useful, but no one is actually trying to exploit it. ....

**Conclusion:** Metadata cannot be readily detected with existing OCR technology, and the IR community is not prepared to address the use of metadata. Therefore, an OCR/IR metadata evaluation conference is not practical.
MATTREX resurrected: NIST Special Database 25
Volume 1
Federal Register Document Image Database

....The database includes scanned images, SGML-tagged ground truth text, commercial OCR results, and image quality assessment results. ....This volume of the database contains 4711 page images scanned binary at 15.75 pixels per millimeter (400 pixels per inch). Cost of the database: $210.00

Standard Reference Data
National Institute of Standards and Technology
100 Bureau Dr., STOP 2310, Gaithersburg, MD 20899-2310
CASIA Database (~1990)
Chinese Academy of Science Institute of Automation

3,755 level-1 set of GB2312-80, 300 writers = 1,126,500

Cheng-Lin Liu:
High Accuracy Handwritten Chinese Character Recognition Using Quadratic Classifiers with Discriminative Feature Extraction. ICPR (2) 2006: 942-945

HCL2000 Database

3,755 level-1 set of GB2312-80, 1000 writers = 3,755,000

H. Zhang, J. Guo, G. Chen, C Li:
A Large-scale Handwritten Chinese Character Data base for Handwritten Character Recognition, ICDAR 2009

TUAT HANDS on-line Databases

~ 3,000,000 patterns in ~4000 Kanji + Symbol categories

S. Jaeger, K. Nakashima, ICDAR 2001

June 9, 2010
2006: Su, Zhang, Guan: HIT-MW

Unconstrained Chinese handwritten characters without preprinted boxes.

853 forms and 186,444 characters

Gray-scale 300dpi BMP

Texts from China Daily
HIT-MW coverage
<table>
<thead>
<tr>
<th>Database</th>
<th>Language</th>
<th>Unit</th>
<th>Year</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highleyman</td>
<td>English</td>
<td>Alphanum</td>
<td>1961</td>
<td>[8]</td>
</tr>
<tr>
<td>Munson</td>
<td>Alphanum</td>
<td></td>
<td>1968</td>
<td>[20]</td>
</tr>
<tr>
<td>Suen</td>
<td>Numeral</td>
<td></td>
<td>1972</td>
<td>[28]</td>
</tr>
<tr>
<td>CENPARMI</td>
<td>Postcode</td>
<td></td>
<td>1992</td>
<td>[30]</td>
</tr>
<tr>
<td>CEDAR</td>
<td>City name</td>
<td></td>
<td>1994</td>
<td>[9]</td>
</tr>
<tr>
<td>CAMBRIDGE</td>
<td>Sentence</td>
<td></td>
<td>1994</td>
<td>[26]</td>
</tr>
<tr>
<td>IAM</td>
<td>Sentence</td>
<td></td>
<td>1998</td>
<td>[16]</td>
</tr>
<tr>
<td>IAAS-4M</td>
<td>Chinese</td>
<td>Character</td>
<td>1985</td>
<td>[15]</td>
</tr>
<tr>
<td>ITRI</td>
<td>Character</td>
<td></td>
<td>1991</td>
<td>[31]</td>
</tr>
<tr>
<td>HCL2000</td>
<td>Character</td>
<td></td>
<td>2000</td>
<td>[36]</td>
</tr>
<tr>
<td>ETL-8</td>
<td>Character</td>
<td></td>
<td>1976</td>
<td>[19]</td>
</tr>
<tr>
<td>ETL-9</td>
<td>Character</td>
<td></td>
<td>1985</td>
<td>[24]</td>
</tr>
<tr>
<td>KU-1</td>
<td>Character</td>
<td></td>
<td>2000</td>
<td>[23]</td>
</tr>
<tr>
<td>IRONOFF</td>
<td>Character</td>
<td></td>
<td>1999</td>
<td>[32]</td>
</tr>
<tr>
<td>GRUHD</td>
<td>Character</td>
<td></td>
<td>2000</td>
<td>[10]</td>
</tr>
<tr>
<td>ISI</td>
<td>Alphnum</td>
<td></td>
<td>2005</td>
<td>[2]</td>
</tr>
</tbody>
</table>
## A taxonomy of test data sets

<table>
<thead>
<tr>
<th>DAS</th>
<th>NON-DAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering drawings</td>
<td>Text (IR, TREK) ?</td>
</tr>
<tr>
<td>Schematic diagrams</td>
<td>Cloud, bubble, &amp; spark chambers</td>
</tr>
<tr>
<td>Maps</td>
<td>Sky (star) pics</td>
</tr>
<tr>
<td>Formulas + equations</td>
<td>Faces</td>
</tr>
<tr>
<td>Signatures, letters, diaries</td>
<td>Finger/palm/foot prints</td>
</tr>
<tr>
<td>Logos</td>
<td>X-rays, CT, MR, PET, ...</td>
</tr>
<tr>
<td>Ballots</td>
<td>Micrographs (cells)</td>
</tr>
<tr>
<td>Illuminated MS, Incunabula, XVI-XIX C books ... ... ...</td>
<td>Plants and flowers</td>
</tr>
<tr>
<td></td>
<td>Features (U C Irvine)</td>
</tr>
</tbody>
</table>

**TABLES**

Forms  

---

June 9, 2010  DAS 2010 (GN)
Four slides from the NSF-III PI Workshop, April 2010

The Scholarly Practice of Information Integration and Informatics

Haym Hirsh
National Science Foundation

Director, Information and Intelligent Systems (CISE)
UC Irvine Machine Learning Repository

- Over 150 Data Sets for Machine Learning and Data Mining
- Median age: 15 years

Should data sets have an expiration date?
Benchmark Data: Reproducible, but Good?

- Benchmark data allow reproducibility
- Reproducibility also depends on software, parameters, etc.
- Benchmark data sets must be representative of the sorts of problems our algorithms will see in practice
- Benchmark data sets must stay timely as technological and scientific advances allow our ambitions to grow
- The pace of data growth makes this difficult
Proprietary Data/Information

- Many innovative ideas involve proprietary or otherwise restricted data/information
- Do we publish innovations that use generally unavailable data/information?
- What can we learn from the medical community?
IAPR TC 5: Benchmarking & Software

Data
A list of publicly available datasets for benchmarking of pattern recognition algorithms

• A list of datasets on the web by datawrangling.com
• UCI standard database in a unified format
• Hand-Written Symbol Recognition
• CAVIAR video sequences
• Sequence Recognition Dataset
• MNist data from Yann LeCun
• UCI Machine Learning Repository
• Youtube 22 Concepts
• USPS data from Max Planck Institute for Biological Cybernetics
• Dataset generator
Machine-print OCR

- APTI: Arabic Printed Text Image Database

Handwriting

On-line

- IAM On-Line Handwriting Database
- UNIPEN database

Off-line

- CEDAR Off-line Handwriting CDROM1
- IAM Database - A full English sentence database for off-line handwriting recognition.
- MARG- Medical Article Records Groundtruth ([1]) is a freely-available repository of document page images and their associated textual and layout data. The data has been reviewed and corrected to establish its "ground truth". Please contact Dr. George Thoma (thoma@lhc.nlm.nih.gov) at the National Library of Medicine for more information.
- Hindi font samples by Andras Kornai, June 5 2003

Miscellaneous Kanji handwritten OCR databases

- IPTP CD-ROM2
comments on test data from www.nagy.virtual.blog
When is $N$ large enough? (Gauss, Chernoff, Guyon, Makhoul, Schwartz, Vapnik)

$$n \approx \left( \frac{z_a}{\beta} \right) \frac{(1-p)}{p} \approx \left( \frac{z_a}{\beta} \right) \frac{1}{p} \approx \frac{2 \ln \alpha}{\beta^2 p} \approx 100$$

So keep sampling until 100 errors have occurred for a 5% ($\alpha$) probability that the true error rate is no more than 20% ($\beta$) greater than the observed value $p$.

Which is the better classifier?
Track exclusive errors $v_1$ and $v_2$ by each classifier. Then the first classifier is probably better if

$$\hat{p}_1 - \hat{p}_2 \geq \frac{z_\alpha}{\sqrt{n}} \sqrt{v_1 - v_2}$$

June 9, 2010
Training on Test Data

• Blatant e.g. estimating language model from a transcript of the test data.
• Subtle e.g. increasing $P(q_\_)$ after finding 38 occurrences of Iraq in the test data.
• Subconscious failing to avert eyes from test data.

If you make more than one run on the test set, report the maximum error!

Testing on Training Data:

![Graph showing error over N for Test set and Training set]
Data Partitioning

• Training – Test (50-50 or 90-10 ?)
• Training – Validation – Test
• Training – Validation 1 – Validation 2 – Validation 3 ... Test
• Leave-one-out, Leave-N-out (cross-validation)
• Non-stationary data
• Sampling with or without replacement

80K samples: 400 sources (writers or fonts), 200 sentences
25% sample: 100 sources, 200 sentences
200 sources, 100 sentences
400 sources, 50 sentences
Representative of what?

• Real data for HP so far primarily from envelopes and checks. Predicting real-world accuracy requires a document census or data from actual operations.

• Copying is different from writing or calculating! Vocabulary, spelling, and grammar are as personal as shape: writer demographics are important (cf. NIST SD3/7)

• In the real world, uniform distributions are rarer than a hen’s tooth! (cf. Benford’s Law for digits, Zipf’s Law for words)

• Ground truth depends on context.
Reporting classifier “performance”

Two-class Confusion Matrix:  5 independent values!
Correct, Error, Reject (C+E+R=1)
Recognition Rate
Precision & Recall (F-score/measure)
Sensitivity & Specificity
Type I & Type II error
False Alarm & Miss
False Positive & False Negative
(+ is good in war, bad in medicine)
Errors of Omission & Commission
Reject-error curve (C.K. Chow 1970)  ROC

<table>
<thead>
<tr>
<th>N=100</th>
<th>Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>True</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>50</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
</tbody>
</table>
Preprocessing considered harmful

Please don’t

binarize,
normalize,
segment,
deskew,
denoise,
thin, fatten,

or otherwise “improve” test data!

Instead, save your algorithm and parameter settings in the database.
Wizard Words mean exactly what you want them to mean:

Concept
Model
Semantics
Context
Ontology

Knowledge
Information
Data
Interpretation
Understanding

as in: “We situate our semantic concept models in an ontological context.”

Un-, non-, semi-supervised;
adaptive, self-adaptive,
teaching, training, learning

Document
Noise

- Real noise is never i.i.d.
- Bits don’t flip by themselves.
- Random-phase noise is unavoidable but benign.
- Bleed-through and leaky margins are infectious.
- Document noise is different from digitization noise.
- A good scanner is worth three noise filters.
- Adding noise to develop robust algorithms is like ear-training with a gong!
- Every $n^{th}$ document should be a calibration target.
Proofreading in context is error prone

This is a review, from an intuitive rather than a mathematical perspective, of the statistical foundations of adaptive recognition systems. Key considerations in adaptive classification are priors, sample size and sampling strategy, labels, statistical dependencies, and dimensionality. The small-sample bias and variance of maximum likelihood, maximum \textit{a posteriori} and Bayes estimators are compared in a small concrete case. Iterative expectation maximization for estimating the sufficient statistics of mixtures is illustrated in a simple setting. It is shown that correlation among features is sometimes unjustly maligned. A counterintuitive increase in the error rate after adding a second feature is traced to the curse of dimensionality. Adaptive classification is presented in the context of both parametric and non-parametric (nearest neighbors and neural nets) estimation. Some recent theoretical results and not-so-recent experimental observations on hybrid classification (based on both labeled and unlabeled samples) are summarized.

Ack: X-ing a Paragraph, Edgar Allan Poe, 1850
Homogenous class display for OCR

11%

June 9, 2010
Foreign is relative

- Language model
- Σχριπτ
- Allographs and variants
- Diacritics (English has only tittles)
- Numerals, punctuation
- Pure or mixed
Keep on testing, testing, testing: progress in document analysis has long been driven by sound experiments on carefully prepared test data.

Thank you